Number: 226513
Acronym: Webdam
Title: Foundations of Web Data Management

ERC Grant

Final Activity Report

Period covered: 01/12/2008 to 01/12/2013
Start date of project: 01/12/2008
PI: Serge Abiteboul
Host: INRIA Saclay – Île-de-France
Date of preparation: 11 December 2013
Duration: 5 years

Original goals The Webdam ERC grant (S. Abiteboul) started in December 2008. The goal is to develop a formal model for Web data management that would open new horizons for the development of the Web in a well-principled way, enhancing its functionality, performance, and reliability. More specifically, the aim is to create a universally accepted formal framework for describing complex and flexible interacting Web applications featuring notably data exchange, sharing, integration, querying and updating. It is also to develop formal foundations enabling peers to concurrently reason about global data management activities and cooperate in solving specific tasks with desired quality of service.

Contents

Chapter 1

Overview

1.1 Major achievements

After 5 years, we are happy to report that the Webdam project has achieved its main goals. After a planning period where we identified research issues, we started investigating the main research themes of Webdam. We introduced major contributions in:

- distributed knowledge bases, notably the Webdamlog language and the system supporting it;
- probabilistic data processing, notably around Probabilistic XML; and
- distributed workflows, from Active XML artifacts to Collaborative workflows.

Among the elements of the visibility of Webdam, one should mention:

- During the 5 years, Serge Abiteboul served as program chair of two of the most prestigious international conferences of the area (VLDB in Lyon and ICDE in Hannover). He was elected member of the French Academy of sciences and of the Academy Europea, as well as fellow of the ACM. He was appointed member of the French Digital Council. Finally, he received the Milner Award from the Royal Society.
- Members of Webdam were quite successful in their careers in academia, notably in Oxford, Tel Aviv and UCSD, or industry.

One should also mention:

Books A textbook on Web data management published at Cambridge University Press [98] and a textbook on Data science (in French) published at Fayard [97]. Both are available freely, respectively at:

- `http://webdam.inria.fr/Jorge`
- `http://webdam.inria.fr/College`
  (English translation at `http://lecons-cdf.revues.org/558`.)

Workshops Webdam organized five Webdam international workshops. Webdam participated in the organization of two International Dagstuhl workshops, one on data-centric workflows [107], and one on distributed data management [106] (together with the FP7 FoX project managed by Luc Segoufin). Webdam coorganized a workshop with the MoDaS ERC Project managed by Tova Milo on Crowd data sourcing [123].

1.2 Publishable brief summary

The Webdam ERC grant was a 5-year project that started in December 2008. The goal was to develop a formal model for Web data management that would open new horizons for the development of the Web in a well-principled way, enhancing its functionality, performance, and reliability.

Information of interest may be found on the Web in a variety of forms, in many systems, with different access protocols. For instance, a standard user may have information on many devices (smartphone, laptop, TV box, etc.), many systems (mailers, blogs, web sites, etc.), many social networks (Facebook, Picasa, etc.). This same user may have access to more information from family, friends, associations, companies, etc., or organizations (tax, health, etc.). The control and management of this diversity are today beyond the skill of casual users. Facing similar issues, companies see the cost of managing and integrating information skyrocketing. The thesis of Webdam is that managing this diversity of data can be achieved using a *distributed knowledge base* handling both data and meta-data, as well as access control and localization information, in a unique holistic setting. We believe that complex Web data management tasks currently requiring deep expertise will be greatly facilitated by the automatic reasoning of the inference engine of the knowledge base. We have obtained fundamental results in that direction and started experimenting with a prototype system.

Distributed knowledge base and Webdamlog As a foundation for managing distribution, we studied a model of a distributed knowledge base, that handles data and meta-data, as well as access control and localization, in a unique integrated setting. The main contribution is a novel rule-based language, namely Webdamlog, featuring the new concept of *delegation*. Using delegation, peers can exchange knowledge and distribute computation. We have implemented a system supporting Webdamlog, studied optimization techniques adapted to that setting, and evaluated the performance of the system, notably in presence of access control.

Imprecise data and Probabilistic XML Data from the Web are imprecise and uncertain. To manage this imprecision in a well-principled way, we have made significant advances in the field of probabilistic databases, and specifically, probabilistic XML. (XML is a semi-structured data model, the standard for data exchange on the Web). We have introduced new tractable probabilistic models for representing uncertain hierarchical information, and carried out in-depth studies of query evaluation, aggregation, and updates in various probabilistic XML models.

Business artifacts and Collaborative workflows Also, when supporting complex activities in a Web setting, one typically has to organize the cooperation between possibly many systems, and notably the sequencing of their tasks. The specification of such sequencing, sometimes referred to as choreography, is little understood. We pursued an original approach that models tasks with pieces of data, that are called *business artifacts* (following IBM terminology). The evolution of an artifact is constrained by rules on the evolution of the data. Using this approach, we developed fundamental works in order to understand the intrinsic nature of workflows shared by collaborative systems.

Webdam has stressed education. In particular,

- A textbook (advanced undergraduate or graduate level) on Web data management has been published at Cambridge University Press [98] in 2009. The book is available for free on the Webdam Web site at http://webdam.inria.fr/Jorge.
- A book "Sciences des données" has been published at Fayard [97] in 2012. The book is available for free at http://lecons-cdf.revues.org/506 and in English translation by Liz Libbrecht at http://lecons-cdf.revues.org/558.

1.3 Major difficulties encountered

Complexity of the model At the early stages of the project, lots of work was devoted to the Active XML model and important results obtained. When considering issues such as trust, access rights, or provenance in the context of social data, the tree aspect of Active XML model turned out to complicate some of the issues. So, we refocused on the relational model and Datalog-style languages for some of the more recent works, e.g., the work around Webdamlog. See Section 3.2.1. Future works should reconcile the two approaches.

Localization In a first phase, the Webdam project was involving researchers of the Institut National de Recherche en Informatique et Automatique (Inria), from the former teams Gemo/Leo at University Paris Sud (now Oak team) and Dahu at Ecole Nationale Supérieure de Cachan (ENS Cachan). The project also rapidly involved researchers at Télécom ParisTech, around Pierre Senellart, notably on probabilistic data. It turned out to be more complicated than expected to focus the group in such a distributed environment. The research was concentrated after a couple of years at ENS Cachan and Télécom ParisTech.

Human Resources These form the main asset of a project such as Webdam. The main reason of the success of Webdam was that we could bring together some incredible talents. This brought some unexpected inherent issue: While we could offer only temporary positions, talented people were offered permanent positions elsewhere (in top places such as Oxford, Tel Aviv, UCSD...) and naturally accepted them. Although this was a difficulty, we also found that this was an opportunity for Webdam to grow by developing collaboration with some of the members who left. Such collaborations are notably on-going with Tel Aviv University and UCSD.

1.4 List of keywords

Distributed data, Database, Data extraction, Data integration.
Data model, Query languages.
Uncertainty, Inconsistency, Probabilistic data.
Semistructured data model, XML, Active XML, Probabilistic XML.
Workflow, Business artifact.
Social Networks, Content recommendation, Corroboration.
World Wide Web, Web searching, Web indexing, Web applications.
Web services, Web languages.
Knowledge, Knowledge representation, Metadata, Datalog.
Webdamlog, Delegation.
Distributed query processing and optimization.

Chapter 2

Project outputs

Number: 226513
Acronym: Webdam
Title: Foundations of Web Data Management
Project output records

2.1 Publications

In the bibliography entries, the keywork *ack* specifies that the corresponding publication acknowledges the ERC funding of Webdam. The keyword *open* specifies that the access to this publication is open followed by a link to the electronic version.

The publications marked open are all reachable from the Webdam publications page.

Journals

[1] Serge Abiteboul, Pierre Bourhis, and Victor Vianu. Comparing workflow specification languages: A matter of views. *ACM Transactions on Database Systems*, 37(2):10, 2012. ack, open `http://webdam.inria.fr/wordpress/wp-content/uploads/2012/12/AbiteboulComparing.pdf`.

[2] Serge Abiteboul, Bogdan Cautis, and Tova Milo. Reasoning about XML update constraints. *Journal of Computer and System Sciences*, 2009. open http://www.math.tau.ac.il/~milo/projects/di_xml_semidata/papers/GemoReport-463.pdf.

[3] Serge Abiteboul, T.-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Capturing continuous data and answering aggregate queries in probabilistic XML. *ACM Transactions on Database Systems*, 36(4):25, 2011. ack, open http://pierre.senellart.com/publications/abiteboul2011capturing.pdf.

[4] Serge Abiteboul, Georg Gottlob, and Marco Manna. Distributed XML design. *Journal of Computer and System Sciences*, 77(6):936–964, 2011. ack, open http://arxiv.org/pdf/1012.2648v1.pdf.

[5] Serge Abiteboul, Yannis Katsis, and Balder Ten Cate. On the equivalence of distributed systems with queries and communication. *Journal of Computer and System Sciences*, 2013. ack, open http://hal.inria.fr/docs/00/87/90/29/PDF/axml-equiv.pdf.

[6] Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, and Pierre Senellart. On the expressiveness of probabilistic XML models. *Journal on Very Large Databases*, 18(5):1041–1064, October 2009. ack, open http://pierre.senellart.com/publications/abiteboul2009expressiveness.pdf.

[7] Serge Abiteboul and Neoklis Polyzotis. Searching shared content in communities with the data ring. *IEEE Data Engineering Bulletin*, 32(2):44–51, 2009. ack, open http://sites.computer.org/debull/A09June/alkis_dataengi1.pdf.

[8] Serge Abiteboul, Luc Segoufin, and Victor Vianu. Modeling and verifying active XML artifacts. *IEEE Data Engineering Bulletin*, 32(3):10–15, 2009. ack, open http://hal.inria.fr/docs/00/42/94/84/PDF/AbiteboulSegoufinVianu09.pdf.

[9] Serge Abiteboul, Luc Segoufin, and Victor Vianu. Static analysis of active XML services. *ACM Transactions on Database Systems*, 34(4), 2009. open http://www.lsv.ens-cachan.fr/~segoufin/Papers/File/axml.pdf.

[10] Bogdan Cautis and Evgeny Kharlamov. Answering queries using views over probabilistic XML: complexity and tractability. *PVLDB*,

5(11):1148–1159, 2012. open http://www.inf.unibz.it/~kharlamov/files/papers/2012/vldb/cautis2012answering.pdf.

[11] Cédric du Mouza, Witold Litwin, and Philippe Rigaux. Large-scale Indexing of Spatial Data in Distributed Repositories: the SD-Rtree. *Journal on Very Large Databases*, 18(4):933–958, 2009. open http://www.springerlink.com/index/0h53343652743r21.pdf.

[12] Clément Genzmer, Volker Hudlet, Hyunjung Park, Daniel Schall, and Pierre Senellart. The SIGMOD 2010 Programming Contest: A Distributed Query Engine. *SIGMOD Record*, 2010. ack, open http://pierre.senellart.com/publications/genzmer2010sigmod.pdf.

[13] Georg Gottlob and Pierre Senellart. Schema Mapping Discovery from Data Instances. *Journal of the ACM*, 2010. ack, open http://hal.inria.fr/inria-00537238/PDF/paper.pdf.

[14] David Gross-Amblard. Query-preserving watermarking of relational databases and XML documents. *ACM Transaction on Database Systems*, 36:3, 2011. open http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.8734.

[15] Wojciech Kazana and Luc Segoufin. First-order query evaluation on structures of bounded degree. *Journal on Logical Methods in Computer Science*, 2011. ack, open http://arxiv.org/abs/1105.3583.

[16] Wojciech Kazana and Luc Segoufin. First-order query evaluation on structures of bounded degree. *Journal of Logical Methods in Computer Science*, 7(2:20), June 2011. ack, open http://arxiv.org/abs/1105.3583.

[17] Julien Lafaye, Jean Béguec, David Gross-Amblard, and Anne Ruas. Blind & squaring-resistant watermarking of vectorial building layers. *GeoInformatica*, 2011. ack.

[18] Stefan Manegold, Ioana Manolescu, Loredana Afanasiev, Jianlin Feng, Gang Gou, Marios Hadjieleftheriou, Stavros Harizopoulos, Panos Kalnis, Konstantinos Karanasos, Dominique Laurent, Mihai Lupu, Nicola Onose, Christopher Ré, Virginie Sans, Pierre Senellart, Tianyi Wu, and Dennis Shasha. Repeatability & workability evaluation of SIGMOD 2009. *SIGMOD Record*, 38(3):40–43, September 2009. open http://pierre.senellart.com/publications/manegold2009repeatability.pdf.

[19] Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. PARIS: Probabilistic alignment of relations, instances, and schema. *PVLDB Journal*, 5(3):157–168, 2011. ack, open http://suchanek.name/work/publications/vldb2012.pdf.

[20] Balder ten Cate, Laura Chiticariu, Phokion Kolaitis, and Wang-Chiew Tan. Laconic Schema Mappings: Computing the Core with SQL Queries. *Journal on Very Large Databases*, 2009. ack, open http://www.vldb.org/pvldb/2/vldb09-595.pdf.

[21] Balder ten Cate, Tadeusz Litak, and Maarten Marx. Complete Axiomatizations of XPath Fragments. *Journal of Applied Logic*, 2009. open http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.141.247&rep=rep1&type=pdf.

Conferences

[22] Serge Abiteboul, Yael Amsterdamer, Daniel Deutch, Tova Milo, and Pierre Senellart. Finding optimal probabilistic generators for XML collections. In *Proceedings of the International Conference on Database Theory*, Berlin, Germany, March 2012. ack, open http://pierre.senellart.com/publications/abiteboul2012finding.pdf.

[23] Serge Abiteboul, Émilien Antoine, Gerome Miklau, Julia Stoyanovich, and Vera Zaychik Moffitt. Introducing access control in webdamlog. *Proceedings of the 14th International Symposium on Database Programming Languages*, 2013. ack, open http://arxiv.org/pdf/1307.8269v1.

[24] Serge Abiteboul, Émilien Antoine, and Julia Stoyanovich. Viewing the Web as a distributed knowledge base. In *Proceedings of the 28th International Conference on Data Engineering*, pages 1–4, 2012. ack, open http://hal.inria.fr/hal-00703210/PDF/11dataengi.pdf.

[25] Serge Abiteboul, Meghyn Bienvenu, Alban Galland, and Émilien Antoine. A rule-based language for Web data management. In *Proceedings of the Symposium on Principles of Database Systems*, 2011. ack, open http://hal.inria.fr/docs/00/58/28/91/PDF/pods17a-abiteboul.pdf.

[26] Serge Abiteboul, Pierre Bourhis, Alban Galland, and Bogdan Marinoiu. The AXML artifact model. In *Proceedings of the International Conference on Temporal Representation and Reason-*

ing, 2009. ack, open http://www.labri.fr/perso/anca/docflow/publications_files/ABGM.pdf.

[27] Serge Abiteboul, Pierre Bourhis, and Bogdan Marinoiu. Efficient maintenance techniques for views over active documents. In *Proceedings of the International Conference on Extending Database Technology*, pages 1076–1087, 2009. open http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.9831&rep=rep1&type=pdf.

[28] Serge Abiteboul, Pierre Bourhis, and Bogdan Marinoiu. Satisfiability and relevance for queries over active documents. In *Proceedings of the Symposium on Principles of Database Systems*, pages 87–96, 2009. open http://leo.saclay.inria.fr/publifiles/gemo/GemoReport-10019.pdf.

[29] Serge Abiteboul, Pierre Bourhis, Anca Muscholl, and Zhilin Wu. Recursive queries on trees and data trees. In *Proceedings of the 16th International Conference on Database Theory*, pages 93–104, 2013. ack, open http://hal.inria.fr/hal-00809297.

[30] Serge Abiteboul, Pierre Bourhis, and Victor Vianu. Comparing Workflow Specification Languages: A Matter of Views. In *Proceedings of the International Conference on Database Theory*, 2011. ack, open http://www.edbt.org/Proceedings/2011-Uppsala/papers/icdt/a9-abiteboul.pdf.

[31] Serge Abiteboul, Pierre Bourhis, and Victor Vianu. Highly expressive query languages for unordered data trees. In *Proceedings of the 15th International Conference on Database Theory*, pages 46–60, 2012. ack, open http://www.edbt.org/Proceedings/2012-Berlin/papers/icdt/a10-Abiteboul.pdf.

[32] Serge Abiteboul, T-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Aggregate Queries for Discrete and Continuous Probabilistic XML. In *Proceedings of the International Conference on Database Theory*, pages 50–61, 2010. ack, open http://hal.inria.fr/inria-00537632/PDF/main.pdf.

[33] Serge Abiteboul, Daniel Deutch, and Victor Vianu. Deduction with Contradictions in Datalog. In *International conference on database theory*, Athenes, Grèce, 2014. ack, open http://hal.inria.fr/hal-00923265.

[34] Serge Abiteboul, Georg Gottlob, and Marco Manna. Distributed XML Design. In *Proceedings of the Symposium on Principles of Database Systems*, pages 247–258, 2009. ack, open `http://fox7.eu/wp-content/uploads/pods2.pdf`.

[35] Serge Abiteboul, Pierre Senellart, and Victor Vianu. The ERC Webdam on foundations of Web data management. In *Proceedings of the International World Wide Web Conference (Companion Volume)*, pages 211–214, 2012. ack, open `http://pierre.senellart.com/publications/abiteboul2012erc.pdf`.

[36] Serge Abiteboul, Balder Ten Cate, and Yannis Katsis. On the Equivalence of Distributed Systems with Queries and Communication. In *Proceedings of the International Conference on Database Theory*, 2011. ack, open `http://www.edbt.org/Proceedings/2011-Uppsala/papers/icdt/a13-abiteboul.pdf`.

[37] Serge Abiteboul and Victor Vianu. Collaborative data-driven workflows: think global, act local. In *Proceedings of the 32nd symposium on Principles of database systems*, pages 91–102, New York, NY, USA, États-Unis, 2013. ACM. ack, open `http://hal.inria.fr/hal-00840306`.

[38] Yael Amsterdamer, Daniel Deutch, Tova Milo, and Val Tannen. On provenance minimization. In *Proceedings of the Symposium on Principles of Database Systems*, pages 141–152, New York, NY, USA, 2011. ACM. ack, open `http://www.cs.bgu.ac.il/~deutchd/publications/pods2011a.pdf`.

[39] Yael Amsterdamer, Daniel Deutch, and Val Tannen. Provenance for aggregate queries. *The Computing Research Repository*, abs/1101.1110, 2011. ack, open `http://arxiv.org/pdf/1101.1110v1`.

[40] Yael Amsterdamer, Yael Grossman, Tova Milo, and Pierre Senellart. Crowd mining. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 241–252, New York, NY, USA, 2013. ACM.

[41] Manuel Atencia, Jérôme Euzenat, Giuseppe Pirrò, and Marie-Christine Rousset. Alignment-based trust for resource finding in semantic p2p networks. In *International Semantic Web Conference*, pages 51–66, 2011. ack, open `http://webdam.inria.fr/wordpress/wp-content/uploads/2011/09/iswc11.pdf`.

[42] Michael Benedikt, Pierre Bourhis, and Pierre Senellart. Monadic datalog containment. In *Proceedings of the 39th International Colloquium on Automata, Languages and Programming*, pages 79–91, 2012. ackopen `http://pierre.senellart.com/publications/benedikt2012monadic.pdf`.

[43] Michael Benedikt, Georg Gottlob, and Pierre Senellart. Determining Relevance of Accesses at Runtime. In *Proceedings of the Symposium on Principles of Database Systems*, 2011. ack, open `http://fox7.eu/wp-content/uploads/pods16d-benedikt.pdf`.

[44] Michael Benedikt, Dan Olteanu, Evgeny Kharlamov, and Pierre Senellart. Probabilistic XML via Markov Chains. In *Proceedings of the International Conference on Very Large Databases*, pages 770–781, 2010. ack, open `http://hal.archives-ouvertes.fr/hal-00537778/PDF/main.pdf`.

[45] Diego Calvanese, Evgeny Kharlamov, Marco Montali, Ario Santoso, and Dmitriy Zheleznyakov. Verification of inconsistency-aware knowledge and action bases. In *Proc. of the 23rd Int. Joint Conf. on Artificial Intelligence (IJCAI 2013)*, 2013. ack, open `http://hal.inria.fr/hal-00817495`.

[46] Diego Calvanese, Evgeny Kharlamov, Werner Nutt, and Dmitriy Zheleznyakov. Evolution of DL-Lite Knowledge Bases. In *Proceedings of International Semantic Web Conference*, pages 112–128, 2010. ack, open `http://hal.archives-ouvertes.fr/hal-00537785/PDF/CalvaneseEtAl2010EvolutionDLLite.pdf`.

[47] Elio Damaggio, Alin Deutsch, and Victor Vianu. Artifact systems with data dependencies and arithmetic. In *Proceedings of the International Conference on Database Theory*, pages 66–77, New York, NY, USA, 2011. ACM. ack, open `http://www.edbt.org/Proceedings/2011-Uppsala/papers/icdt/a8-damaggio.pdf`.

[48] Cédric du Mouza, Witold Litwin, Philippe Rigaux, and Thomas Schwarz. AS-Index: A Structure For String Search Using n-grams and Algebraic Signatures. In *Proceedings of the International Conference on Information and Knowledge Management*, 2009. ack, open `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.161.2287&rep=rep1&type=pdf`.

[49] Zoé Faget, Philippe Rigaux, David Gross-Amblard, and V. Thion Goasdoué. Modeling synchronized time series. In *Proceedings of the International Database Engineering and Applications Symposium*, pages 82–89, 2010. ack, open http://cedric.cnam.fr/~thionv/Publications/2010/IDEAS2010_Faget_Gross_Rigaux_Thion.pdf.

[50] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating Information from Disagreeing Views. In *Proceedings of the International Conference on Web Search and Data Mining*, 2010. ack, open http://hal.inria.fr/inria-00429546/PDF/document.pdf.

[51] Bernardo Cuenca Grau, Ernesto Jiménez-Ruiz, Evgeny Kharlamov, and Dmitriy Zheleznyakov. Ontology evolution under semantic constraints. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference*, 2012. ack-open http://www.inf.unibz.it/~kharlamov/files/papers/2012/kr/grau2012ontology.pdf.

[52] David Gross-Amblard, Philippe Rigaux, Lylia Abrouk, and Nadine Cullot. Fingering Watermarking in Symbolic Digital Scores. In *Proceedings of the International Conference on Music Information Retrieval*, 2009. open http://www.lamsade.dauphine.fr/scripts/FILES/publi1101.pdf.

[53] Wojciech Kazana and Luc Segoufin. Enumeration of first-order queries on classes of structures with bounded expansion. In *Proceedings of the Symposium on Principles of Database Systems*, pages 297–308, 2013. ack, open http://www.lsv.ens-cachan.fr/Publis/PAPERS/PDF/KS-lmcs11.pdf.

[54] Evgeny Kharlamov and Dmitriy Zheleznyakov. Capturing instance level ontology evolution for dl-lite. In *International Semantic Web Conference*, pages 321–337, 2011. open http://www.inf.unibz.it/~kharlamov/files/papers/2011/iswc/KharlamovZheleznyakov2011Capturing.pdf.

[55] Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. A theory of pricing private data. In *ICDT*, pages 33–44, 2013. open http://arxiv.org/pdf/1208.5258v1.

[56] Chao Li and Gerome Miklau. Optimal error of query sets under the differentially-private matrix mechanism. In *Proceedings of the 16th*

International Conference on Database Theory, ICDT '13, pages 272–283, New York, NY, USA, 2013. open http://arxiv.org/pdf/1202.3399v3.

[57] Bruno Marnette and Floris Geerts. Static analysis of schema-mappings ensuring oblivious termination. In *Proceedings of the International Conference on Database Theory*, 2010. ack, open http://disi.unitn.it/~p2p/RelatedWork/Matching/icdt10_Marnette.pdf.

[58] Bruno Marnette, Giansalvatore Mecca, and Paolo Papotti. Scalable data exchange with functional dependencies. In *Proceedings of the International Conference on Very Large Databases*, 2010. ack, open http://www.vldb.org/pvldb/vldb2010/pvldb_vol3/R09.pdf.

[59] Asma Souihli and Pierre Senellart. Optimizing approximations of dnf query lineage in probabilistic xml. In *Proceedings of the 29th International Conference on Data Engineering*, pages 721–732, 2013. ack, open http://pierre.senellart.com/publications/souihli2013optimizing.pdf.

[60] Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. PARIS: Probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment*, December 2011. ack, open http://suchanek.name/work/publications/vldb2012.pdf.

[61] Fabian M. Suchanek, David Gross-Amblard, and Serge Abiteboul. Watermarking for ontologies. In *ISWC 2011: Proceedings of the 10th International Semantic Web Conference*, 2011. ack, open http://suchanek.name/work/publications/iswc2011.pdf.

[62] Balder ten Cate and Gaelle Fontaine. An Easy Completeness Proof for the Modal mu-calculus on Finite Trees. In *Proceedings of the International Workshop on Fixed Points in Computer Science*, 2009. ack, open http://users.soe.ucsc.edu/~gaelle/fossacsend2.pdf.

[63] Remi Tournaire, Alexandre Termier, Jean-Marc Petit, and Marie-Christine Rousset. Combining logic and probabilities for discovering mappings between taxonomies. In *Proceedings of the International Conference on Knowledge Science, Engineering and Management*, 2010. ack, open http://www.springerlink.com/content/d110g2t62n5gq8mm/.

[64] Victor Vianu. Automatic verification of database-driven systems: a new frontier. In *Proceedings of the International Conference on Database*

Theory, pages 1–13, 2009. open http://www.edbt.org/Proceedings/2009-StPetersburg/icdt/papers/p0001-Vianu.pdf.

Workshops

[65] Serge Abiteboul, Meghyn Bienvenu, and Daniel Deutch. Deduction in the presence of distribution and contradictions. In *Proceedings of the 15th International Workshop on the Web and Databases*, pages 31–36, 2012. ack, open DeductioninthePresenceofDistributionandContradictions.

[66] Serge Abiteboul, Meghyn Bienvenu, Alban Galland, and Marie-Christine Rousset. Distributed Datalog Revisited. In *Proceedings of the Datalog 2.0 International Workshop*, 2011. ack, open http://hal.inria.fr/inria-00540814/PDF/datalog20-serge.pdf.

[67] Serge Abiteboul, Meghyn Bienvenu, Alban Galland, and Marie-Christine Rousset. Distributed Datalog Revisited. In *Datalog 2.0 Workshop*, Oxford, United Kingdom, 2011. ack, open http://hal.inria.fr/docs/00/54/21/17/PDF/datalog20-serge.pdf.

[68] Serge Abiteboul, Alban Galland, and Neoklis Polyzotis. A model for web information management with access control. In *Proceedings of the International Workshop on the Web and Databases*, 2011. ack, open webdb2011.rutgers.edu/papers/Paper%2013/WebDB11.pdf.

[69] Loredana Afanasiev and Balder ten Cate. On Core XPath with Inflationary Fixed Points. In *Proceedings of the International Workshop on Fixed Points in Computer Science*, page 11, 2009. ack, open http://cs.ioc.ee/fics09/fics09proc.pdf.

[70] Yael Amsterdamer, Daniel Deutch, and Tova Milo. On the optimality of top-k algorithms for interactive web applications. In *Proceedings of the International Workshop on the Web and Databases*. ACM, 2011. ack, open http://webdb2011.rutgers.edu/papers/Paper%2011/InteractiveWebApp.pdf.

[71] Diego Calvanese, Evgeny Kharlamov, Werner Nutt, and Dmitriy Zheleznyakov. Updating ABoxes in DL-Lite. In *Proceedings of Alberto Mendelzon Workshop on Foundations of Data Management (AMW)*, 2010. ack, open http://hal.archives-ouvertes.fr/hal-00537787/PDF/main.pdf.

[72] Bogdan Cautis and Evgeny Kharlamov. Challenges for View-Based Query Answering over Probabilistic XML. In *Proceedings of the Alberto Mendelzon International Workshop on Foundations of Data Management (AMW)*, 2011. ack, open http://hal.inria.fr/inria-00591913/PDF/main.pdf.

[73] A. Gheerbrant and Balder ten Cate. Craig Interpolation for Linear Temporal Languages. In *Proceedings of International Workshop on Computer Science Logic*. Springer, 2009. ack, open http://www.springerlink.com/content/l01324k014410278/.

[74] Bernardo Cuenca Grau, Evgeny Kharlamov, and Dmitriy Zheleznyakov. Ontology contraction: Beyond the propositional paradise. In *Proceedings of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management*, pages 62–74, 2012. ack, open http://www.inf.unibz.it/~kharlamov/files/papers/2012/amw/grau2012paradise.pdf.

[75] Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Updating Probabilistic XML. In *Proceedings of Workshop on Updates in XML; at the International Conference on Extending Database Technology*, 2010. ack, open http://hal.archives-ouvertes.fr/hal-00537793/PDF/main.pdf.

[76] Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Value Joins are Expensive over (Probabilistic) XML. In *Proceedings of the Workshop on Logic in Databases ; at the International Conference on Extending Database Technology*, 2011. ack, open http://hal.inria.fr/inria-00591905/PDF/main.pdf.

[77] Evgeny Kharlamov and Dmitriy Zheleznyakov. On prototypes for Winslett's semantics of DL-Lite ABox evolution. In *Proc. Description Logics Workshop (DL)*, 2011. ack, open http://www.inf.unibz.it/~kharlamov/files/papers/2011/dl/kharlamov2011prototypes.pdf.

[78] Evgeny Kharlamov and Dmitriy Zheleznyakov. Understanding Inexpressibility of Model-Based ABox Evolution in DL-Lite. In *Proceedings of the Alberto Mendelzon International Workshop on Foundations of Data Management*, 2011. ack, open http://hal.inria.fr/inria-00591918/PDF/main.pdf.

[79] Marilena Oita, Antoine Amarilli, and Pierre Senellart. Cross-fertilizing deep Web analysis and ontology enrichment. In *Proceedings of the Very Large Data Search*, pages 5–8, 2012. ack,open http://pierre.senellart.com/publications/oita2012crossfertilizing.pdf.

[80] Marilena Oita and Pierre Senellart. Archiving Data Objects using Web Feeds. In *Proceedings of the International Workshop on Web Archiving*, 2010. ack, open http://hal.inria.fr/inria-00537962/PDF/iwawienna.pdf.

[81] Marilena Oita and Pierre Senellart. Deriving Dynamics of Web Pages: A Survey. In *Proceedings of the Temporal Workshop on Web Archiving*, 2011. ack, open http://hal.inria.fr/inria-00588715/PDF/survey.pdf.

[82] Fabian Suchanek, Alessandro Bozzon, Emanuele Della Valle, Alessandro Campi, and Stefania Ronchi. Towards an Ontological Representation of Services in Search Computing. In Stefano Ceri and Marco Brambilla, editors, *New Trends in Search Computing*. Springer, 2011. ack, open http://hal.inria.fr/inria-00591790/PDF/seco_yago.pdf.

[83] Remi Tournaire, Alexandre Termier, Jean-Marc Petit, and Marie-Christine Rousset. Probamap: a scalable tool for discovering probabilistic mappings between taxonomies. In *Proceedings of the International Workshop on Automated Knowledge Base Construction*, 2010. open akbc.xrce.xerox.com/IMG/pdf/CameraReady_TournaireR.pdf.

[84] Dmitriy Zheleznyakov, Diego Calvanese, Evgeny Kharlamov, and Werner Nutt. Updating TBoxes in DL-Lite. In *Proceedings of International Workshop on Description Logics*, 2010. ack, open http://hal.archives-ouvertes.fr/hal-00537790/PDF/main.pdf.

Demonstrations (in proceedings of international conferences)

[85] Serge Abiteboul, Yael Amsterdamer, Tova Milo, and Pierre Senellart. Auto-completion learning for XML. In *Proceedings of the 2013 ACM SIGMOD Special Interest Group on Management Of Data*, pages 669–672, 2012.

[86] Serge Abiteboul, Émilien Antoine, Gerome Miklau, Julia Stoyanovich, and Jules Testard. Rule-Based Application Development using Webdamlog. In *SIGMOD - Proceedings of the 2013 ACM SIGMOD Special*

Interest Group on Management Of Data, New York, United States, 2013. hal.inria.fr/hal-00817791.

[87] Serge Abiteboul, Émilien Antoine, Gerome Miklau, Julia Stoyanovich, and Jules Testard. Rule-Based Application Development using Webdamlog. In *BDA - La 29e édition des journées Bases de Données Avancées*, Nantes, France, 2013.

[88] Serge Abiteboul, Pierre Bourhis, Bogdan Marinoiu, and Alban Galland. [Demo] AXART - Enabling Collaborative Work with AXML Artifacts. In *Proceedings of the International Conference on Very Large Database*, 2010.

[89] Serge Abiteboul, Ohad Greenshpan, Tova Milo, and Neoklis Polyzotis. Matchup: Autocompletion for mashups (demo). In *Proceedings of the International Conference on Database Engineering*, pages 1479–1482, 2009.

[90] Yael Amsterdamer, Yael Grossman, Tova Milo, and Pierre Senellart. Crowd miner: Mining association rules from the crowd. In *Proceedings of the Very Large Data Bases International Conference*, VLDB '13, 2013.

[91] Émilien Antoine, Alban Galland, Kristian Lyngbaek, Amélie Marian, and Neoklis Polyzotis. [Demo] Social Networking on top of the WebdamExchange System. In *Proceedings of the International Conference on Data Engineering*, 2011.

[92] Michael Benedikt, Tim Furche, Andreas Savvides, and Pierre Senellart. ProFoUnd: program-analysis-based form understanding. In *Proceedings of the International World Wide Web Conference*, pages 313–316. ACM, 2012.

[93] Luying Chen, Michael Benedikt, and Evgeny Kharlamov. QUASAR: querying annotation, structure, and reasoning. In *Proceedings of the 15th International Conference on Extending Database Technology*, pages 618–621. ACM, 2012.

[94] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, Edwin Lewis Kelham, Gerard De Melo, and Gerhard Weikum. [Demo] YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. In *Proceedings of the World Wide Web Conference - Demos*, 2011.

[95] Asma Souihli and Pierre Senellart. Demonstrating ProApproX 2.0: a predictive query engine for probabilistic XML. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2734–2736, 2012.

[96] Fabian M. Suchanek and David Gross-Amblard. Adding fake facts to ontologies. In *Proceedings of the International World Wide Web Conference*, pages 421–424, 2012.

Book

[97] Serge Abiteboul. *Sciences des données: de la Logique du premier ordre à la Toile.* Fayard, 2012. ack open `http://abiteboul.com/College/lecon.htm`.

[98] Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. *Web Data Management.* Cambridge University Press, 2011. ack open `http://webdam.inria.fr/Jorge`.

Book Chapter

[99] Serge Abiteboul, Omar Benjelloun, and Tova Milo. Active XML. In *Encyclopedia of Database Systems*, pages 38–41. Springer, 2009.

[100] Michael Benedikt and Pierre Senellart. Databases. In Edward K. Blum and Alfred V. Aho, editors, *Computer Science. The Hardware, Software and Heart of It.* Springer-Verlag, January 2012.

[101] Evgeny Kharlamov and Pierre Senellart. Modeling, Querying, and Mining Uncertain XML Data. In Andrea Tagarelli, editor, *XML Data Mining: Models, Methods, and Applications.* IGI Global, 2011. ack.

[102] Benny Kimelfeld and Pierre Senellart. Probabilistic XML: Models and complexity. In *Advances in Probabilistic Databases for Uncertain Information Management*, pages 39–66. Springer, 2013.

Others

[103] Serge Abiteboul. Sharing distributed knowledge on the Web (invited talk). In *CSL*, pages 6–8, 2012.

[104] Serge Abiteboul. Viewing the web as a distributed knowledge base. In *Description Logics*, 2012.

[105] Serge Abiteboul, Klemens Böhm, Christoph Koch, and Kian-Lee Tan, editors. *Proceedings of the International Conference on Data Engineering*. IEEE Computer Society, 2011.

[106] Serge Abiteboul, Alin Deutsch, Thomas Schwentick, and Luc Segoufin, editors. *Proceedings of the Dagstuhl Seminar 11421 "Foundations of distributed data management"*, 2011, to appear.

[107] Serge Abiteboul, Agnes Koschmider, Andreas Oberweis, and Jianwen Su, editors. *Proceedings of the Dagstuhl Seminar 10151 "Enabling Holistic Approaches to Business Process Lifecycle Management"*, 2010.

[108] Serge Abiteboul, Volker Markl, Tova Milo, and Jignesh Patel. Special issue: Best papers of international conference on very large databases 2009. *Journal on Very Large Databases*, 20(2):155–156, 2011.

[109] Serge Abiteboul, Volker Markl, Tova Milo, Jignesh Patel, and Philippe Rigaux, editors. *Proceedings of the 27th International Conference on Very Large Databases*, 2009.

[110] Émilien Antoine. *Distributed data management with the rule-based language: Webdamlog.* PhD thesis, Université Paris-Sud, 2013.

[111] Pierre Bourhis. *On the Dynamics of Active Documents for Distributed Data Management.* PhD thesis, Université Paris-Sud, 2011.

[112] Alban Galland. *Distributed Data Management with Access Control.* PhD thesis, Université Paris-Sud, 2011.

[113] David Gross-Amblard. *Tatouage des bases de données.* Habilitation à Diriger des Recherche, Université de Bourgogne, 12 2010.

[114] Wojciech Kazana. *Query Evaluation with Constant Delay.* PhD thesis, Laboratoire Spécification et Vérification, ENS Cachan, France, 2013.

[115] Evgeny Kharlamov. *A Probabilistic Approach to XML Data Management.* PhD thesis, Faculty of Computer Science, Free University of Bozen-Bolzano, 2011.

[116] Hady Lauw, Ralf Schenkel, Fabian Suchanek, Martin Theobald, and Gerhard Weikum. Harvesting Knowledge from Web Data and Text. In *Proceedings of the International Conference on Information and Knowledge Management*, 2010.

[117] Bogdan Marinoiu. *Analysis and verification of distributed systems*. PhD thesis, Université Paris Sud, 2009.

[118] Marilena Oita. *Deriving Semantic Objects from the Structured Web*. PhD thesis, Télécom ParisTech, October 2012.

[119] Nicoleta Preda, Gjergji Kasneci, Fabian Suchanek, Thomas Neumann, Wenjun Yuan, and Gerhard Weikum. Active Knowledge: Dynamically Enriching RDF Knowledge Bases by Web Services. In *Proceedings of the ACM SIGMOD Conference on the Management of Data*, 2010. Tutorial.

[120] Pierre Senellart. *Probabilistic XML: A Data Model for the Web*, June 2012. Habilitation to supervise research, Université Pierre et Marie Curie.

[121] Fabian Suchanek, Martin Theobald, Gerhard Weikum, Hady Lauw, and Ralf Schenkel. Semantic Knowledge Bases from Web Sources. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2011. Tutorial.

[122] Fabian Suchanek, Aparna Varde, Richi Nayak, and Pierre Senellart. The Hidden Web, XML and Semantic Web: A Scientific Data Management Perspective. In *Proceedings of the International Conference on Extending Database Technology*, 2011. ack. Tutorial.

[123] WebDam-MoDas workshop. `http://www.cs.tau.ac.il/workshop/modas/`, 2012. Eilat, Israel.

2.2 Research expeditions

Not applicable

2.3 Awards and recognitions

2.3.1 Research prizes

- S. Abiteboul was the recipient of the Milner's Award of the Royal Society. ("This award is made for outstanding achievement in computer science by a European researcher.")

- S. Abiteboul was Chair Professor at Collège de France 2011-2012.

- S. Abiteboul was Francqui Chair Professor at Université de Namur 2012-2013.

- Victor Vianu received the 2010 Alberto O. Mendelzon Test of Time Award of ACM PODS for the article *Typing for XML Transformers*, joint with Tova Milo and Dan Suciu.

- Daniel Deutch won the ICDT 2011 Best Student Paper Award

- Meghyn Bienvenu received the Association Francaise pour l'Intelligence Artificielle Prize for her thesis in 2010.

- Fabian Suchanek received the Otto-Hahn-Medal, the Dissertation Award of the Max Planck Society in 2010.

- Fabian Suchanek was also selected for a research group leader stipend by the Max Planck Society in 2010.

- Fabian Suchanek won the ACM Dissertation Award Honorable Mention for his dissertation in 2010.

- The demo of the YAGO ontology [94], to which Fabian Suchanek contributed, won the Best Demo Award at the WWW 2011 conference.

2.3.2 Memberships of a learned society

- Serge Abiteboul has been elected in 2009 member of the French Academy of Sciences.

- S. Abiteboul has been elected in 2011 member of the Academia Europa.

- Marie-Christine Rousset has been elected in 2011 senior member of the Institut Universitaire de France.

- Serge Abiteboul has been elected in 2012 Fellow of the Association of Computing Machinary (ACM).

- Victor Vianu hes been elected on 2013 Fellow of the American Association for the Advancement of Science (AAAS).

2.3.3 Memberships of editorial boards

- Victor Vianu is Editor-in-Chief of the *Journal of the ACM* and Area Editor for *ACM Trans. on Computational Logic* (logical aspects of databases).
- Serge Abiteboul was a member 2010-2012 of the steering committee of the *Proceedings of the VLDB Endowment* (*PVLDB*) journal, a recently launched journal.
- Pierre Senellart is Information Director of *JACM*.

2.4 Patents, licencing, intellectual property

We licenced the software used in two demonstrations [88, 91].

2.5 Dissemination to non-academic audience

S. Abiteboul participated to a popularization book on mathematics, "Mathématiques, l'explosion continue", with an article, Chercher sur le Web : juste un point fixe et quelques algorithmes.

S.Abiteboul's radio talk shows: Science Publique (France Culture), Place de La Toile (France Culture), Autour de la question (RFI). He participated in a radio program for the 20 years of the Web: Le Téléphone Sonne, Alain Bedouet, France Inter.

Marie-Christine Rousset participated in *Xavier de La Porte* podcast about "Genèse d'un algorithm" related to the paper on interstices.info also in french.

Serge Abiteboul and Pierre Senellart wrote an article on "Un déluge de données", in Pour la science sur le Big bang numérique, 2013.

Presentations to a general audience:

- The Future of Communications and the Web, Futur en Seine, 2011.
- Le Web de demain: promesses et écueil, Conférences de la Communauté Israélite de Genève, 2010.
- Demain, la République du Web : une utopie ? La Cantine, 2009.
- Web data management, Colloque Une histoire de DIM, Domaines d'Intérêt Majeur, Paris 2009.
- Web et Industrie, Rencontres INRIA-Industrie, 2009.

- Gestions de données distribuées, Perspectives IT pour le Codir des Clubs OLG, 2009.

- Données, Information, connaissances : challenges, Data Excellence Paris Conference, Paris 2012

- La gestion de données à l'heure de la Toile, Data Tuesday, (avec Fernando Velez), Issy-les-Moulineaux 2012

- Quelle cuisine pour les données du Web ? Let's imagine the Future workshop, Rennes 2012

- Les connaissances du Web, Entretiens de la Cité, Lyon, 2013.

- Les connaissances de la toile, Conférence cultures numériques, éducation aux médias et à l'information, Lyon, 2013.

- Des données, à l'information, aux connaissances : le Web de demain, Conférences Science et société, Nancy, 2013.

- La vie dans le milieu académique et Comment choisir un sujet de thèse, Congrès de la Société informatique de France, Nice, 2013.

- Vers une nouvelle science du risque, Journée SCOR sur les Big data et les assurrances, Paris (même sujet aux Journéee Ifpas de l'assurance), 2013.

- A propos du rapport de l'Académie des Sciences "L'enseignement de l'informatique en France - Il est urgent de ne plus attendre", Journées Educatec-Educatice, Paris, 2013.

- Open data et santé, Congrés Health IT, Systèmes d'Information en Santé, Paris, 2013.

- La Toile des Fictions, Séminaire Vérifiction, CNAM, Paris, 2013.

- Des données à l'information, aux connaissances : le web de demain Forum régionaux des Savoirs, Rouen, 2013.

- Web et Connaissances, Journéee Economie de la connaissance et économie numérique : l'innovation en question, IHEST, Saclay, 2013.

- S. Abiteboul a été également auditionné à l'Assemblée nationale par la commission des affaires économiques, Mission d'information sur l'économie numérique; et par l'Office parlementaire d'évaluation des choix scientifiques et technologiques, sur le Risque numérique. 2013.

S.Abiteboul's interviews in the press:

- Big data: Les utilisations, qui concernent surtout le commerce, pourraient s'étendre au domaine social, Le Monde, Octobre 2013, Interview par Sophy Caulier
- Les algorithmes: un jeu d'enfant, Famille chrétienne, Septembre 2013, Interview par Jacques Henno
- Recherche profs d'informatique désespérément, 01Net, avec Gilles Dowek et Colin de la Higuera, 2013.
- Construisons un Web des savoirs, Le Monde, 2012
- Le Web redéfinit sans cesse les échanges d'information, Le Figaro 2012
- L'enseignement de l'informatique en classes prépas, 01.Net (avec Colin de La Higuerra)
- Le Big Data est avant tout un effet de mode, O1Net
- Sur les liens entre labos publics et universitaires, 01.Net 2012.
- L'informatique est une science bien trop sérieuse pour être laissée aux informaticiens, Le monde.fr (avec Colin de la Higuera et Gilles Dowek)
- L'important sur Internet, c'est de trouver la bonne information, Lepoint.fr
- Pour la Science, 2009.
- Le Nouvel Economiste, 2009,
- L'informaticien, 2009
- Le Sévrien 2009
- Science et Vie, Nos 3 questions, 2009
- Specif, 2009

2.6 Dissemination

2.6.1 PhD theses

Bogdan Marinoiu (2009) *Analysis and verification of distributed systems* [117]

Pierre Bourhis (2011) *On the dynamics of active documents for distributed data management* [111].

Evgeny Kharlamov (2011) *A Probabilistic Approach to XML Data Management* [115]

Alban Galland (2011) *Distributed Data Management with Access Control* [112]

Marilena Oita (2012) *Deriving Semantic Objects from the Structured Web* [118]

Émilien Antoine (2013) *Distributed data management with the rule-based language: Webdamlog* [110]

Wojciech Kazana (2013) *Query Evaluation with Constant Delay* [114]

2.6.2 Habilitation theses

David Gross-Amblard (2010) *Database Watermarking* [113]

Pierre Senellart (2012) *Probabilistic XML: A Data Model for The Web* [120]

2.6.3 Participation in conference programs

Members of Webdam were regularly invited to serve in the program committees of the best conferences.

- Serge Abiteboul served as General Program Chair of[1] the International Conference on Data Engineering (ICDE) 2011 in Hanover, Germany (one of the main conferences on database systems) [105].

  He served as General Program Chair of the International Conference on Very Large Databases (VLDB) 2009, the main conference on database systems) in Lyon, France [109, 108].

  These are two out of the three top conferences in database systems. The third one is the ACM SIGMOD International Conference on the

[1] In this text, the first time a conference or journal is encountered, we give the name in full and the acronym. We then use the acronym.

Management of Data (SIGMOD) where Webdam was also strongly represented; see Attachment 2.6.

- Ioana Manolescu served as Co-Chair of the Web Engineering Track in the World Wide Web Conference (WWW) 2009. She served as Co-Chair of the "Systems, Experiments, Applications" Track in ICDE 2011.

- Pierre Senellart was Submission Chair of VLDB 2009, Tutorial Co-Chair of ICDE 2010, and Program Chair of the Industrial Track of the International Conference on Extending Database Technology (EDBT) 2011.

- Serge Abiteboul has served on the Program Committees of International Conference: notably, on Business Process Modelling 2010, on Database theory (ICDT) 2012, World Wide Web (WWW) 2008 et 2012, on Exchanging DataBase Technology (EDBT) 2014, on Principles of Database Systems (PODS) 2014.

- Meghyn Bienvenu has served on the Program Committees of the International Joint Conference on Artificial Intelligence (IJCAI) 2011, the AAAI Conference on Artificial Intelligence 2010 and 2011, and the European Conference of Artificial Intelligence (ECAI) 2010. She is also a member of the Program Committee of the International Description Logic Workshop (2009, 2010, 2011), in addition to serving on the workshop's Steering Committee (elected 2009-2012).

- David Gross-Amblard has served on the Program Committees of Bases de Données Avancées (BDA 2010 & 2011 & 2012 (Demo) & 2013).

- Ioana Manolescu served on the Program Committee of SIGMOD 2009.

- Ioana Manolescu and Pierre Senellart participated in the Repeatability & Workability Evaluation Track of SIGMOD 2009 [18].

- Pierre Senellart served on the Program Committees of the ACM Symposium on PODS 2009 & 2012, BDA 2010, 2012, the Conference on Information and Knowledge Management (CIKM) 2010, ICDE 2010, ICDT 2011, WWW 2009 & 2010, EDBT 2013 (demo), SIGMOD 2013, and VLDB 2013.

- Fabian Suchanek served on the Program Committee of WWW 2011.

- Victor Vianu served on the Program Committees of ACM Symposium on PODS 2009 & 2012, VLDB 2009, IEEE Symposium on Logic in

Computer Science (LICS) 2010, VLDB 2010, International Symposium on Foundations of Information and Knowledge Systems (FoIKS) 2010, Alberto Mendelzon Workshop on Foundations of Data Management (AMW) 2010 & 2011, and the Datalog 2.0 Workshop 2012.

2.6.4 Participation in conference organization

Webdam participated in the organization of *two major workshops* centered around its research themes both in the Leibniz Center for Informatics (Dagstuhl):

Serge Abiteboul co-organized with Andreas Oberweis (KIT, Germany) and Jianwen Su (UCSB, USA) the Dagstuhl Workshop on Enabling Holistic Approaches to Business Process Lifecycle Management (04/2010) [107]

Serge Abiteboul and Luc Segoufin are co-organizing with Alin Deutsch (UCSD, USA) and Thomas Schwentick (TU Dortmund, Germany) the Dagstuhl Workshop on Foundations of Distributed Data Management (10/2011) [106]

Serge Abiteboul co-organized with Tova Milo (Tel Aviv) the WebDam-MoDaS Workshop on Web data management and Crowdsourcing, Eilat, Israel 2012. Émilien Antoine co-organized the brainstorming session at that WebDam-MoDaS workshop.

Pierre Senellart organized (together with Serge Abiteboul) the SIGMOD 2010 Programming Contest. Teams of contestants from degree-granting institutions had to develop an efficient distributed query engine on top of an in-memory index. The competition received much attention, with 29 teams from 23 different institutions worldwide [12].

Fabian Suchanek organized the PhD Workshop at CIKM 2011, 2012, 2013, together with Anisoara Nica (Sybase Inc.). The workshop aims to give PhD students an opportunity to present their dissertation research at a relatively early stage. The workshop focuses on databases, information retrieval and knowledge management.

2.6.5 Webdam workshops

The following International Webdam events were organized:

- A Webdam kickoff meeting at ENS Cachan in January 2009.

- A brainstorming workshop at Télécom ParisTech, Paris 2009 (Post VLDB) with members of the advisory board[2].

- A Workshop on Modal Logic (organized by Balder ten Cate) at ENS Cachan 2009.

- A mid-term Webdam Workshop at Télécom ParisTech, Paris 2011.

- Serge Abiteboul and P. Senellart co-organized the Webdam "Data in the Wild" Workshop, Paris 2012.

- Serge Abiteboul, Pierre Senellart and Victor Vianu organized the *Final Workshop for Web data management* (nicknamed Webdone) in Paris 2013.

2.6.6 Invited presentations and tutorials

- Serge Abiteboul the following presentations:

  - A keynote presentation on "Active XML Artifact" at the International Symposium on Temporal Representation and Reasoning (TIME) 2009 [26]. He gave a keynote presentation on "Web Information Management and Knowledge Bases" at the 10th International Conference on Web Engineering, Vienna (07/2010)
  - An invited presentation on "Web Data Management" at the Datalog 2.0 Workshop, held in March 2010, at Oxford University [66]
  - An invited presentation on "Object Databases" at the Dagstuhl workshop on Relationships, Objects, Roles, and Queries in Modern Programming Languages (04/2010)
  - An invited presentation on "Workflow Specification Languages for Active Documents" at the Fox Workshop in Amsterdam (05/2010)
  - An invited presentation on "WebdamExchange: A Model for Data Access on the Web" at the Workshop on Formal Methods for Web Data Trust and Security
  - An invited presentation on "Web Data Management" at Centre d'Alembert and French Academy of Sciences.

[2] This informal board consists of Prof. Peter Buneman, University of Edinburgh, Prof. Stefano Ceri, Politecnico di Milano, Prof. Georg Gottlob, Oxford University, Prof. Rick Hull, IBM Watson Research Center, Prof. Tova Milo (Chair), Tel Aviv University, Prof. Moshe Vardi, Rice University.

- A panel on “Distributed Data Management” at ICDE2011.
- Sharing Distributed Knowledge on the Web, Conference on Computer Science Logic, Fontainebleau 2012
- Viewing the Web as a Distributed Knowledge Base, International Workshop on Description Logic, Roma 2012
- Web data management, INTIMATE workshop on “Big Data in Digital Life”, Paris 2012
- Science of data: from first order logic to the Web, Colloque « Translittératies : enjeux de citoyenneté et de créativité » ENS-Cachan et Université Sorbonne nouvelle, Cachan 2012
- Collective question answering, MSR-INRIA Workshop, Cambridge 2012
- Overview of Webdam, WebDam-MoDaS Workshop, Eilat, Israel 2012
- How can humans and systems collaborate in a social network to answer queries: issues and challenges, panel with P. Buneman, M. Franklin, H.V. Jagadish
- Viewing the Web as a Distributed Knowledge Base, French-Israeli Workshop on Foundations of Computer Science, Paris 2012
- Viewing the Web as a Distributed Knowledge Base, EPFL, Lausanne 2012
- From data and information to knowledge: the Web of tomorrow, Milner’s Lecture, Royal Society, London
- Introducing Access Control in Webdamlog, International Workshop on Database Programming Languages, Trento
- From data and information to knowledge: the Web of tomorrow, Plateforme Intelligence Artificielle, Lille.
- Rule-Based Application Development using Webdamlog, Serge Abiteboul, Conférence sur les Bases de Données Avancées, Nantes

- Serge Abiteboul and Victor Vianu gave two of the invited presentations (out of five) at the PODS special workshop organized to celebrate the 30th anniversary of the conference (06/2011). They spoke respectively on “Trees, semistructured data, and other strange ways to go beyond tables” and on “Database Theory: Back to the Future”.

- Ioana Manolescu gave an invited presentation at XML Symposium & Database and Programming Languages workshops (VLDB’09).

- Pierre Senellart gave a tutorial on distributed computing and indexing at the French-German summer school for young researchers in Frauenchiemsee in July 2011. He was a keynote speaker at the ACM SoICT 2013 conference.

- Pierre Senellart and Fabian Suchanek presented a tutorial on “Data Management over the Deep Web and the Semantic Web” at EDBT 2011 [122].

- Pierre Senellart and Evgeny Kharlamov gave a tutorial at the French database conference on probabilistic data management in Rabat in October 2011.

- Fabian Suchanek was an invited expert at the Search Computing Workshop in Milan. He contributed to a tutorial on “Harvesting the Web of Data” at CIKM 2010, and on “Web-scale Knowledge Bases” at IJCAI 2011.

2.6.7 Education

Abiteboul published “Sciences des données”, Fayard [97] 2012. The book is available on the Web at http://lecons-cdf.revues.org/506 and in English translation by Liz Libbrecht at http://lecons-cdf.revues.org/558.

Serge Abiteboul’s teaching:

- As Francqui chair Professeur at Université de Namur 2012-2013, S. Abiteboul taught a course on Web data Management.

- Serge Abiteboul has been professor at College de France 2011-2012 2012. He organized a 10 hours course on Web data management. He also organized a seminar on the topic with for guests: Moshe Vardi, Anastasia Ailamaki, François Bancilhon, Julien Masanès, Victor Vianu, Tova Milo, Georg Gottlob, Gerhard Weikum, Marie-Christine Rousset, Pierre Senellart.

- School “Imagine the Future in ICT”, organized by ICDT lab. Two courses: (i) Data sciences; (ii) Web search engines.

- Relational databases, undergratuate course, ENS Cachan and ENS Paris - 5 years.

- Web data management, graduate course, MPRI Paris S. Abiteboul, P. Senellart, P. Rigaux - 3 years.

- A course on “Web Data Management” at the 4th Winter School in Hot Topics in Distributed Computing (La Plagne 2011).
- Half day tutorial at the thematic school of the University ES-Sénia, Oran, Algeria and the Journées Nationales APMEP.
- Life in Academia and How to Choose a Thesis Topic, First AVSE Doctoral Workshop, Cachan 2012

2.6.8 Journal editing

- Victor Vianu: Editor-in-chief of JACM, Area editor for ACM Trans. on Computational Logic (logical aspects of databases), Editor of the Database Theory Column of SIGACT News
- Pierre Senellart is Information Director of the *Journal of the ACM*.

2.6.9 Books

We already mentioned the textbook on Data science (in French) published at Fayard [97]. We detail next the book on Web data management that is the core topic of Webdam:

- S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset and P. Senellart have written a textbook entitled *Web Data Management and Distribution*, to be published by Cambridge University Press in 2011 [98]. It covers the recent advances in the modeling, querying, integration and indexing of very large data sets. The focus is on Web scale data management, and the text deals with some of the most important issues that arise in this context (e.g., heterogeneity, volume, distribution in large networks, etc.). The target audience consists of graduate and PhD students, engineers and practitioners seeking an in-depth presentation of the languages, techniques and tools required to build large-scale and distributed information systems. Parts of the book served as material for various courses and events (in France, Hong Kong, Rome, Beijing. . .), including a summer school co-organized by Webdam in Les Houches in May 2010.

2.6.10 Responsibilities

S. Abiteboul is the principal investigator of the European Research Council Grant Webdam on Web Data Management.

S. Abiteboul is chairman of the Scientific Council of Société d'Informatique de France, elected in 2012.

As a member of the Sciences Academy, S. Abiteboul wrote a report on "L'enseignement de l'informatique en France – Il est urgent de ne plus attendre".

S. Abiteboul is since 2013 a member of the Conseil national de la recherche. As a member, he participated in 2013 in reports on Net neutrality, Computer science education, and digital inclusion.

S. Abiteboul is chairman of the INRIA Awards committee.

S. Abiteboul is chairman of the Scientific Board of Société d'Informatique de France.

S. Abiteboul is a member of the Academic Senat of the University Paris-Saclay.

S. Abiteboul is a member of the Academia Europea.

Chapter 3

Attachments

List of attachments:

- Attachment 3.1: Participants
- Attachment 3.2: Research results in more detail

3.1 Participants

In this section, we detail the human participation in Webdam with the date of arrival and departure.

Researchers and professors

- Serge Abiteboul, senior researcher *principal investigator*
- Balder ten Cate, researcher [04/2009–10/2009] [12/2008]
- Daniel Deutch, post doc [02/2011–10/2011]
- David Gross-Amblard, assistant professor at University of Bourgogne, seconding [09/2010–08/2011]
- Yannis Katsis, post doc [09/2009–07/2011]
- Gerome Miklau, associate professor at University of Massachusetts, Amherst [09/2012–06/2013]
- Bruno Marnette, post doc [09/2010–07/2011]
- Philippe Rigaux, professor at Dauphine University, seconding at Webdam [09/2009–10/2010]

- Marie-Christine Rousset, professor at University of Grenoble, seconding at Webdam [09/2009–10/2010]
- Fabian Suchanek, post doc [06/2010–07/2011]
- Victor Vianu, professor at UC San Diego [07/2010–07/2011]&[07/2013-12/2013]

PhD Students

- Yael Amsterdamer, Tel Aviv University [02/2011–10/2011]
- Émilien Antoine, Paris Sud University [10/2010–12/2013]
- Pierre Bourhis, Paris Sud University [12/2008–03/2011]
- Alban Galland, Paris Sud University [12/2008–09/2011]
- Wojciech Kazana, ENS Cachan [02/2010–06/2013]
- Evgeny Kharlamov, Free University of Bozen-Bolzano [01/2009–12/2010]
- Bogdan Marinoiu, Paris Sud University [12/2008–09/2009]
- Marilena Oita, Télécom ParisTech [10/2009–10/2012]

Other full-time members

- Antoine Amarilli, intern (Télécom ParisTech) [04/2012–08/2012]
- Alin Gabriel Tilea, engineer [09/2009–06/2010]
- Kristian Lyngbaek, intern (Delft University) [07/2010–05/2010]
- Marilena Oita, intern [03/2009–07/2009]
- Vanya Petrova, intern [04/2009–07/2009]
- Jules Testart, intern (McGill University) [05/2011–08/2011][09/2012-12/2012]

Collaborators

- Meghyn Bienvenu, researcher at Université Paris Sud [2011–2012]
- Ioana Manolescu, INRIA senior researcher, manager of the Oak team, INRIA Saclay [2008–2009]
- Philippe Rigaux, professor at Paris CNAM [2011–2012]
- Luc Segoufin, INRIA senior researcher, manager of the Dahu team, INRIA Saclay [entire project]
- Pierre Senellart, professor at Télécom ParisTech [entire project]
- Julia Stoyanovich, assistant professor at Drexel University [2011–2013]
- Victor Vianu, professor at UC San Diego [entire project]

Webdam Alumni

Webdam alumni are very successful in academia:

- Antoine Amarilli went to Télécom ParisTech to do a PhD.
- Yael Amsterdamer went to Tel Aviv University to do a PhD.
- Émilien Antoine went to Kyoto Sangyo University as a Postdoc.
- Pierre Bourhis is CNRS Researcher in Lille.
- Balder ten Cate is Assistant Professor at University of California Santa, Cruz.
- Daniel Deutch is Assistant Professor at Tel Aviv University, Israel.
- David Gross-Amblard was promoted to Full Professor at Rennes University.
- Yannis Katsis went to University of California, San Diego as a Postdoc.
- Wojciech Kazana went to University of California, San Diego as a Postdoc.
- Evgeny Kharlamov is Senior Researcher at the University of Oxford.
- Gerome Miklau was promoted to Full Professor at the University of Massachusetts, Amherst.

- Philippe Rigaux was promoted to Full Professor at CNAM, Paris.
- Pierre Senellart was promoted to Full Professor at Télécom ParisTech.
- Fabian Suchanek is Assistant Professor at Télécom ParisTech.
- Balder ten Cate is Assistant Professor at UC Santa Cruz.

Webdam alumni are also successful in other positions:

- Bogdan Marinoiu is engineer at SAP.
- Alban Galland holds a management position at the French "Ministère de l'Économie, des Finances et de l'Industrie".
- Bruno Marnette is engineer at Palantir.
- Marilena Oita is engineer at Internet Memory.
- Philippe Rigaux is consulting at Internet Memory.
- Alin Gabriel Tilea is engineer at Sefas Innovation.

Visitors (one month or more)

- Sihem Amer-Yahia, Yahoo Research [12/2009]
- Balder ten Cate, professor at UC Santa Cruz [04/2009-10/2009]
- Diego Calvanese, associate professor at Free University of Bozen-Bolzano [02/2009]
- Yukiko Kawai, associate professor at Kyoto Sangyo University [04/2013-09/2013]
- Bruno Marnette, PhD student at Oxford [07/2013-08/2013]
- Amélie Marian, assistant professor at Rutgers University [09/2009 and 12/2009]
- Gerome Miklau, associate professor at the University of Massachusetts, Amherst [06/2010]
- Tova Milo, professor at Tel Aviv U. [10/2009]
- Werner Nutt, full professor at Free University of Bozen-Bolzano [04/2009-09/2009]

- Yannis Papakonstantinou, professor at UC San Diego [10/2009]
- Alkis Polizotis, professor at UC Santa Cruz [07/2009]

3.2 Research results in more detail

In this section, we detail the scientific contributions. Most of the cited papers may be found at: `http://webdam.inria.fr/`

3.2.1 The Web as a distributed knowledge base

The management of data in a heavily distributed setting such as the Web poses a number of challenging issues such as how to find data, how to control data access, and how to integrate data coming from different sources. Webdam is investigating a holistic approach to support these complex data management tasks which is based on reasoning over a *distributed knowledge base.*

WebdamExchange We propose a knowledge-base model, called WebdamExchange, for sharing information on the Web, where the information is hosted on different machines that may use different access control and distribution schemes. The model uses logical statements for specifying data, access control, distribution and knowledge about other peers. The statements can be communicated, replicated, queried, and updated, while keeping track of time and provenance. This unified basis allows applications to reason about which data is accessible, where it resides, and how to retrieve it securely.

A system supporting this model has been implemented and demonstrated in the ICDE 2011 conference [91]. The demonstration illustrates how users can keep control over their data even in a social network that facilitates exchanges. In particular, it shows how users within very different data-distribution schemes (centralized, DHT, gossiping in an unstructured P2P, etc.) and different access control schemes, can transparently collaborate while keeping a good control over their own data even when some of their data resides in standard website or social network systems such as Facebook. The demonstration also illustrates how users can even control data using a device with limited storage and processing capabilities such as a smart phone. This research is part of the PhD thesis of Alban Galand under the supervision of Serge Abiteboul.

Webdamlog We are developing the system Webdamlog [24, 103, 104] to address the challenges faced by everyday Web users, who interact with inherently heterogeneous and distributed information. Managing such data is currently beyond the skills of casual users. In Webdamlog, we see the Web as a knowledge base consisting of distributed logical facts and rules. The objective is to enable automated reasoning over this knowledge base, ultimately improving the quality of service and of data. The system supports the Webdamlog language, a Datalog style language with rule delegation.

In WebdamExchange, the peers' policies are hard-wired (coded in Java, in the prototype system). One would like users to be able to specify their own policies or customize existing ones. For that, we want to use a declarative language. Indeed, we strongly believe that the specification of modern distributed applications should be based on a reasonably simple (declarative) language that would hide most of the unnecessary details of distributed data management. This is in the spirit of work from Berkeley University on declarative programming for distributed systems, notably around the Dedalus language.

Preliminary work on this topic was presented at the Datalog 2.0 workshop [66]. Then in [25], we introduced a novel Datalog-style rule-based language, called Webdamlog. In this language, peers exchange messages (i.e. logical facts) as well as rules. The model is formally defined, and its interest for distributed data management is illustrated through a variety of examples. We validate the semantics of our model by showing that under certain natural conditions, our semantics converges to the same semantics as the centralized system with the same rules. Indeed, we can show this is even true when updates are considered. Another major contribution of this work is a study of the impact on expressiveness of 'delegations" (the installation of rules by a peer in some other peer) and explicit timestamps.

Finally, we proposed an implementation of the Webdamlog engine and a framework to develop applications in Webdamlog (demonstrated at SIGMOD 2013) [86]. We implement optimization techniques based on provenance for an efficient evaluation of Webdamlog programs. We also lead a user study to validate that Webdamlog is a declarative that can be understood and written by non-programmers. This research is part of the PhD thesis of Émilien Antoine under the supervision of Serge Abiteboul.

Collaborative Access Control in WebdamLog The management of Web users' personal information is increasingly distributed across a broad array of applications and systems, including online social networks and cloud-based services. While users wish to share and integrate data using these systems, it is increasingly difficult to avoid the risks of unintended disclosures or unauthorized access by applications.

We proposed in [23] a novel access control model that operates within a distributed data management framework based on datalog. Using this model, users can control access to data they own and control applications they run. They can conveniently specify access control policies providing flexible tuple-level control derived using provenance information. We present a formal specification of the model, a theoretical analysis, and an implementation. We show that the computational cost of access control is acceptable. We started considering access right issues in Webdamlog. This is related to specifying access right on views in standard databases. There is also the issues of controlling rules that are run locally but were specified by other peers. This leads to the development of a framework for distributed access control in Webdamlog with the participation of Gerome Miklau from U. Massachusetts (one year visit in Webdam) and Julia Stoyanovich from Drexel University.

3.2.2 The Web as a probabilistic world

The information found on the Web is typically uncertain, imprecise, possibly inconsistent. Also we may wrongly interpret it. This leads to issues such as data quality or trust. This is a major theme for Webdam. Our approach is based on probabilities and more precisely, on *probabilistic trees* (i.e., probabilistic XML). We have made important progress in this topic, studying the modeling, querying, and more generally management, of probabilistic XML.

Models for Probabilistic XML We started by proposing a general model for probabilistic trees that encompasses all previously proposed models (including local and global probabilistic dependencies) and studying the respective expressiveness of these [6]. With this framework, it is possible to represent in a possibly compact manner all discrete and finite probability distributions over trees. This naturally leads to the question of representing continuous distributions, or discrete infinite distributions (with trees of unbounded size). In [32, 3], we provide a formal model for continuous distributions in probabilistic trees and shows that they essentially do not add any complexity to querying tasks, as long as a number of operations (convex sums, differentiation, integration, convolution) can be tractably performed over these distributions, either symbolically or numerically. In [44], a joint work with the FoX FP7 project, we overcome the limitation that trees described in probabilistic XML have a bounded size with the help of recursive Markov chains (RMCs), a formalism for probabilistic processes with recursive calls. We explain how to use them to generate probabilistic trees. We show that some natural restrictions of RMCs (that are strictly more expressive than existing local probabilistic XML models) yield tractable query answering, in some cases assuming unit-cost arithmetics.

Querying Probabilistic XML Previous work (notably by Cohen, Kimelfeld, and Sagiv) had investigated the complexity of, and algorithms for, evaluating tree-pattern queries (and monadic second-order logic) over classical probabilistic XML models. In addition to extending these results to our more general settings (e.g., in [44]), we have extended query capabilities over probabilistic XML in a number of directions. In [76], we extend tree-pattern queries with value joins and show that for simple classes of such queries, there is a dichotomy between tractable and intractable queries over probabilistic XML models with local dependencies. In [32], we provide a systematic study of the complexity of aggregate queries defined with tree-pattern queries with joins (and restrictions of that query language). We show that when only local probabilistic dependencies are considered, some particular aggregate functions (namely, *monoid* ones) are tractable. In [72, 10], we investigate the problem of answering queries using views for probabilistic XML.

In another, more applied, line of work, we investigated efficient approximation of probabilistic XML query results. Because the most appropriate approximation algorithm to use will depend on both the query and the data, we developed in [59] an approach inspired by traditional query optimization to choose an optimal approximation plan for the probability of a query result. The system was also demonstrated in in [95].

Managing Probabilistic XML We looked at other issues of interest pertaining to probabilistic XML. Initial studies of updates in probabilistic XML [6] suggested that the more complex the probabilistic dependencies in data are, the simpler it is to represent the result of a probabilistic update in that model. We investigate the problem further in [75] and show the situation is more contrasted: it is actually possible to represent tractably the result of insertions defined by simple queries in a local dependency model, while this is not possible in a model with dependencies expressed by arbitrary conjunctions. Deletions almost always lead to a combinatorial explosion, except when arbitrary propositional formulas are allowed as probabilistic conditions on nodes of a tree. Finally, in [101], we show how some data mining problems can be solved using probabilistic XML querying techniques.

Inferring Probabilisting XML In [22], we study the problem of, given a corpus of XML documents and its schema, finding an optimal (generative) probabilistic model, where optimality here means maximizing the like- lihood of the particular corpus to be generated. Focusing first on the structure of documents, we present an efficient algorithm for finding the best generative probabilistic model, in the absence of constraints. We further study the problem in the presence of integrity constraints, namely key, inclusion, and domain constraints. We study in this case two different kinds of generators. First, we consider a continuation-test generator that performs, while generating documents, tests of schema satisfiability ; these tests prevent from generating a document violating the constraints but, as we will see, they are computationally expensive. We also study a restart generator that may generate an invalid document and, when this is the case, restarts and tries again. Finally, we consider the injection of data values into the structure, to obtain a full XML document. We study different approaches for generating these values.

Based on these results, we developed ALEX, an Auto-completion Learning Editor for XML [85]. The editor assists the users by providing intelligent auto-completion suggestions.

Most of the works on probabilistic XML developed in the framework of Webdam are sumarized in Pierre Senellart's habilitation thesis [120], and in a survey of the field, published in [102].

Corroboration In an environment with many independent participants, one typically finds conflicting opinions. In [50], we study the problem of corroborating information coming from a large number of participants. We propose and evaluate various algorithms towards this goal that provide improvements over known techniques such as voting, that often already perform reasonably well in practice.

Uncertainty in Crowd Data Sourcing In collaboration with the ERC MoDaS, we developed a framework for mining association rules from the crowd [40]; this requires proper modeling of the uncertainty in what we know about a given association rule given the crowd's current answers. The system developed for crowd mining was also demonstrated in [90].

Deduction in uncertain worlds Motivated by reasoning in distributed environments in which disagreements arise between different actors, we study in [65] deduction (captured by datalog programs) in the presence of inconsistencies (induced by functional dependency (FD) violations). We adopt an operational semantics for datalog with FDs based on inferring facts one at a time, while never violating the FDs. This yields a set of possible worlds that we capture by c-tables of possibly exponential size. We propose to use probabilities to measure this nondeterminism and define a probabilistic semantics that can be captured by probabilistic conditional tables. Not surprisingly, we show that computing the probability of a query answer in our setting is expensive, which leads us to introduce a sampling algorithm to estimate answer probabilities. We then turn our attention to the problem of explaining why a particular answer holds. This leads us to consider two novel notions: the most influential extensional facts, and the most likely proofs for an answer. We study algorithms for ranking facts and proofs based on their contribution to the derivation of an answer. Finally, we consider how our framework can be adapted to a distributed setting, and in particular, how sampling can be performed in a distributed manner.

3.2.3 Languages on trees

We studied in [31] highly expressive query languages for unordered data trees, using as formal vehicles Active XML and extensions of languages in the while family. All languages may be seen as adding some form of control on top of a set of basic pattern queries. The results highlight the impact and interplay of different factors: the expressive power of basic queries, the embedding of computation into data (as in Active XML), and the use of deterministic vs. nondeterministic control. All languages are Turing complete, but not necessarily query complete in the sense of Chandra and Harel. Indeed, we show that some combinations of features yield serious limitations, analogous to FOk definability in the relational context. On the other hand, the limitations come with benefits such as the existence of powerful normal forms. Other languages are “almost” complete, but fall short because of subtle limitations reminiscent of the copy elimination problem in object databases.

In [42], we reconsider the problem of containment of monadic datalog (MDL) queries in unions of conjunctive queries (UCQs). Prior work has dealt with special cases, but has left the precise complexity characterization open. We begin by establishing a 2EXPTIME lower bound on the MDL/UCQ containment problem, resolving an open problem from the early 90’s. We then present a general approach for getting tighter bounds on the complexity, based on analysis of the number of mappings of queries into tree-like instances. We use the machinery to present an important case of the MDL/UCQ containment problem that is in co-NEXPTIME, and a case that is in EXPTIME. We then show that the technique can be used to get a new tight upper bound for containment of tree automata in UCQs. We show that the new MDL/UCQ upper bounds are tight.

3.2.4 Sequencing tasks on the Web

The orchestration of distributed activities (under the responsibility of a conductor) and their choreography (when they are fully autonomous) are complex issues that are essential for a wide range of applications including notably, e-commerce systems, business processes, health-care and scientific workflows. The difficulty is to guarantee consistency or more generally, quality of service, and to statically verify critical properties of the system. This is a theme where Webdam is very active with pioneering works and fundamental results.

There has recently been a proliferation of workflow specification formalisms. One such approach, proposed at IBM by Nigam and Caswell in 2003, is that of *data-centric workflow*, that places data at the center of the process, controlling sequencing by constraining the evolution of this data. This is also the philosophy of Webdam, with evolving trees (Active XML) at the center of the workflow.

In brief, Active XML consists of XML documents (the standard of the Web for data exchange) with embedded function calls. The state of a document evolves depending on the result of internal function calls (local computations) or external ones (interactions with users or other services). Functions can be naturally used to model tasks in a workflow. They return documents that may be active, so may in turn activate new sub-tasks, thus having the ability to naturally specify a hierarchy of tasks.

Verifying temporal properties of runs In [9, 8], we study the verification of temporal properties of runs of Active XML systems, specified in a tree-pattern-based temporal logic, namely Tree-LTL, expressing a rich class of semantic properties of the application. The main results establish the boundary of decidability and the complexity of automatic verification of Tree-LTL properties.

In collaboration with UC San Diego and IBM organized around Prof. Victor Vianu, we studied the verification problem for a restricted class of data-centric workflows (in the spirit of IBM's Business Artifacts). However, these early results suffer from an important limitation: they fail in the presence of even very simple data dependencies or arithmetic, both crucial to real-life business processes. In [30], we extend the artifact model and verification results to alleviate this limitation. We identify a practically significant class of business artifacts with data dependencies and arithmetic, for which verification is decidable.

Alternative specification approaches There are different approaches to workflow specification such as automata-based, logic-based, or predicate-based control of function calls. In [30], we reconcile them by showing that they can be seen as closely related points of view on the evolution of data.

Specifically, we propose in [30] a flexible framework for comparing workflow specification languages, in which the pertinent aspects to be taken into account are defined by *views*. We use these views to compare the expressiveness of different workflow specification mechanisms based on automata, pre/post conditions, and temporal constraints. (Note the intrinsic difficulty of the problem because of the lack of a standard yardstick for measuring expressiveness.)

The Active XML Artifact model In [26], we introduce the *Active XML artifact model* to capture data and workflow management activities in distributed settings. We argue that the model is a natural extension of the Business Artifact model of Nigam and Caswell.

A prototype named AXART based on this model has been developed and demonstrated at VLDB 2010 [88]. It uses an application taken from the movie industry, that specifies task sequencing when managing actors applications for roles in films. Because the system builds on [26], it allows for complex organization of tasks. Because it borrows from [30], the system supports different ways of expressing workflow constraints, a rather unique feature.

This research is part of the PhD thesis of Pierre Bourhis, under the direction of Serge Abiteboul.

3.2.5 Other topics

Analysis of distributed data management

With the evolution of the Web and the emergence of universal standards for data exchange, data management is becoming increasingly distributed. As a first step in setting theoretical foundations for distributed data management systems, we study the equivalence of such systems [36]. To model these systems, we use the Active XML framework. The resulting model is expressive enough to capture distributed systems that are recursive, utilize asynchronous communication and operate on data streams.

As our model is quite general, the equivalence problem is undecidable. We exhibit restrictions of the model, in terms of query languages used in the function calls and the presence or absence of external function calls, for which equivalence can be effectively decided. We study the computational complexity of the equivalence problem, and for a limited class of distributed systems present a full axiomatization of equivalence.

Ontologies

Ontology alignment The data on the Web is not limited to HTML pages. Nowadays, an increasing part of the Web is taken by the *Semantic Web*. The Semantic Web can be seen as a gigantic, distributed and interlinked entity-relationship graph. It contains machine-interpretable information from encyclopedias, gazetteers, government censuses, music databases, libraries and numerous other sources. Each such sub-graph is commonly known as an *ontology*.

Our work in this area progresses on four research avenues. First, we work on contributing data to the Semantic Web. Together with researchers from the Max Planck Institute in Germany, we are developing techniques to enrich the YAGO ontology [1] by a temporal and spatial dimension [94]. Second, we investigate the interaction of ontologies and Web services. Together with researchers from Stefano Ceri's group in Milan, we are working on bridging the gap between the Semantic Web and the Web service technology developed in the Search Computing project [82]. Third, we are working on algorithms that can detect and merge equivalent entities, classes and properties in different ontologies. With more and more ontologies becoming available on the Web, it is becoming important to discover their overlap in order to harness their synergies. Last, we are working on ownership proofs for ontological data, in the spirit of [14].

One of the main challenges that the Semantic Web faces is the integration of a growing number of independently designed ontologies. In [19] work, we present PARIS, an approach for the automatic alignment of ontologies. PARIS aligns not only instances, but also relations and classes. Alignments at the instance level cross-fertilize with alignments at the schema level. Thereby, our system provides a truly holistic solution to the problem of ontology alignment. The heart of the approach is probabilistic, i.e., we measure degrees of matchings based on probability estimates. This allows PARIS to run without any parameter tuning. We demonstrate the efficiency of the algorithm and its precision through extensive experiments. In particular, we obtain a precision of around 90 % in experiments with some of the world's largest ontologies. PARIS has also been successfully applied to information extraction tasks, in the setting of deep Web querying [79].

[1] See article by F.M. Suchanek and G. Kasneci and G. Weikum in WWW 2007.

Evolution of ontologies Description Logics (DLs) provide excellent mechanisms for representing structured knowledge by means of ontologies, and as such they constitute the foundations for the various fragments of OWL, the standard ontology language of the Semantic Web. As a response to the dynamicity of the Web, we have studied the problem of evolution for ontologies expressed in different DLs. We propose some fundamental principles that evolution should respect in [46], and study evolution for ontologies expressed in OWL 2 QL, a tractable fragment of OWL 2. We review known model and formula-based approaches for evolution and exhibited their limitations. Building upon the insights gained, we propose two novel formula-based approaches and develop polynomial time algorithms for them [46, 71, 84]. We also consider in [78] model-based approaches to analyze why these approaches raise difficulties when considering ontology evolution.

Watermarking Webdam has also addressed the problem of evolution of ontologies, in the presence of semantic constraints [51], and of ontology watermarking either through deletion [61] or addition [96] of facts.

Web archiving

In the course of the Webdam project, we investigated how actual data from the Web can be archived in a timely manner, going further than what is currently done by archival crawlers. In particular, we study how Web feeds can be used for archiving Web pages containing temporal data objects, such as blog posts or news items. We use RSS or Atom feeds to extract these Web objects and to detect change in the context of an incremental crawl. To detect change on crawled Web pages that have a Web feed associated, we propose an algorithm that extracts the information of interest (the data object), with the aim of analyzing changes effectively, without being tricked by possible changes in the surrounding boilerplate. See [80] for further details, and [118] for an extension to the detection of Web objects through prominent keywords found on the page, without any need of RSS feeds. We also present in [81] a survey on techniques used for timestamping and detecting changes of pages on the Web.

Deep Web data

A part of the Web is hidden to traditional Web search engines, because its data lie behind Web form interfaces. Understanding how to exploit in an automatic manner data from the deep Web is a challenging task, and we have laid some foundations of deep Web data management in Webdam. In [13], we propose a formal model for discovering the optimal schema mapping between two deep Web sources, allowing to understand how the data presented in one source can be derived from the other. In [43], we investigate the notion of *relevance* of a form interface to a query, i.e., when using a particular form can be useful in a query evaluation task. This led to the problem of discovering access limitations in real-world interfaces, through static analysis of client-side code accompanying Web forms [92].

This work is in collaboration with the ERC Diadem project (G. Gottlob, University of Oxford).

Query enumeration

In many applications, the output of a query may have a huge size and enumerating all the answers may already consume too much of the allowed resources. When this is the case, it may be appropriate to first output a small subset of the answers and then, on demand, output a subsequent small number of answers and so on until all possible answers have been exhausted. To make this even more useful, it is preferable to be able to minimize the time necessary to output the first answers as well as the time between consecutive sets of answers; this second time interval is known as the *delay*. This problem can be tackled by computing appropriate auxiliary index structures. The goal is to construct in reasonable time (say in linear time in the size of the database), index structures allowing extremely short delay in the enumeration process (say constant time). We are interested in understanding under what assumptions (on both the query language and the nature of the database) such index and algorithms exist. This approach was shown possible in [15] within the following scenario: solutions to any query expressible in first-order logic over a relational structure of bounded degree can be enumerated, after a linear preprocessing, with a constant delay.

This research is part of the PhD thesis of Wojciech Kazana, under the direction of Luc Segoufin.

Monitoring

We have worked on the conception and implementation of tools for monitoring Peer to Peer Systems. Before Webdam started, a system named P2PMonitor has been developed for this purpose. It is a P2P system itself, with peers exchanging messages by Web-service calls. In Webdam, we have studied what we believe is at the heart of such distributed monitoring: the maintenance of views over active documents. Indeed, the monitoring problem can be seen as aggregating streams into an active document and incrementally evaluating a tree-pattern query over this active document. In [27], we develop algorithmic Datalog-based foundations for such an incremental query processing. We also study theoretical issues raised in this context, such as whether a query over some active documents can be satisfied and whether a particular input stream is relevant for a given query [28].

This research is part of the PhD theses of Bogdan Marinoiu and Pierre Bourhis, both under the direction of Serge Abiteboul.

Distributed XML design

A distributed XML document is an XML document that spans several machines or Web repositories. We assume that a distribution design of the document tree is given, providing an XML tree some of whose leaves are "docking points", to which XML subtrees can be attached. These subtrees may be provided and controlled by peers at remote locations, or may correspond to the result of function calls, e.g., Web-services. If a global type τ, e.g. a DTD, is specified for a distributed document T, it would be most desirable to be able to break this type into a collection of local types, called a local typing, such that the document satisfies τ if and only if each peer (or function) satisfies its local type. In [34], we lay out the fundamentals of a theory of local typing and provide formal definitions of three main variants of locality: local typing, maximal local typing, and perfect typing, the latter being the most desirable.